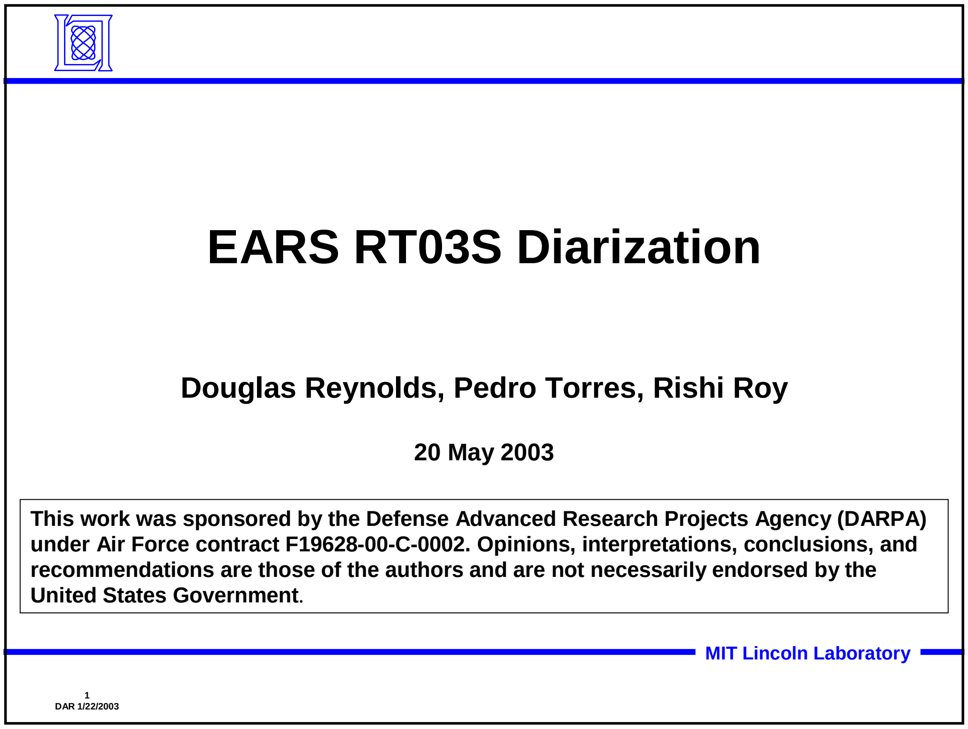
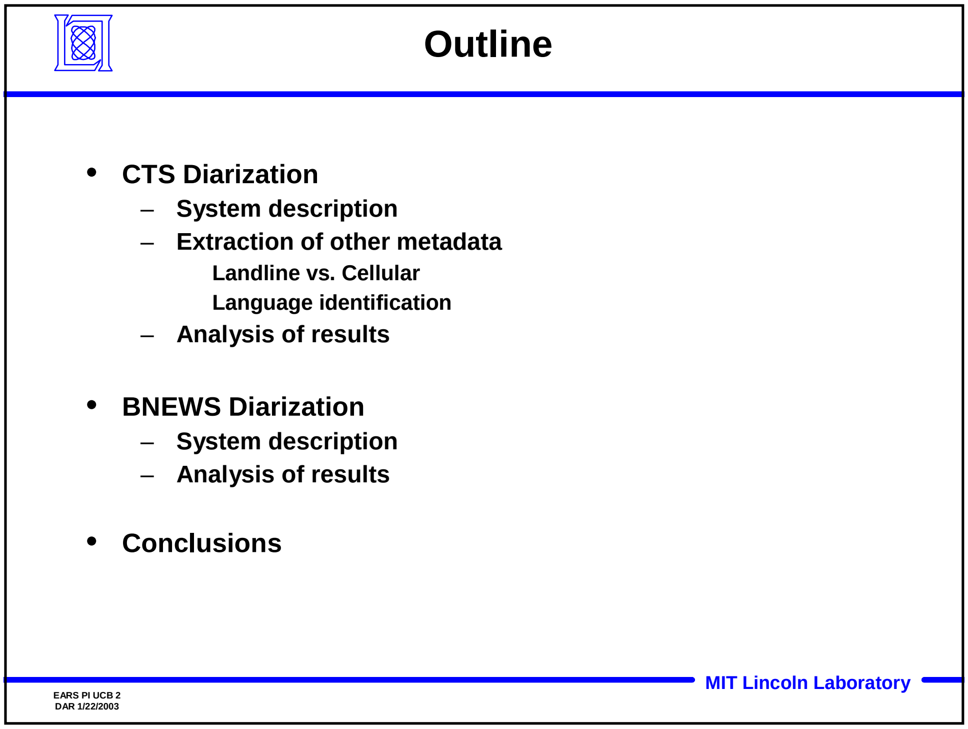


EARS RT03S Diarization

Douglas Reynolds, Pedro Torres, Rishi Roy

20 May 2003

This work was sponsored by the Defense Advanced Research Projects Agency (DARPA) under Air Force contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

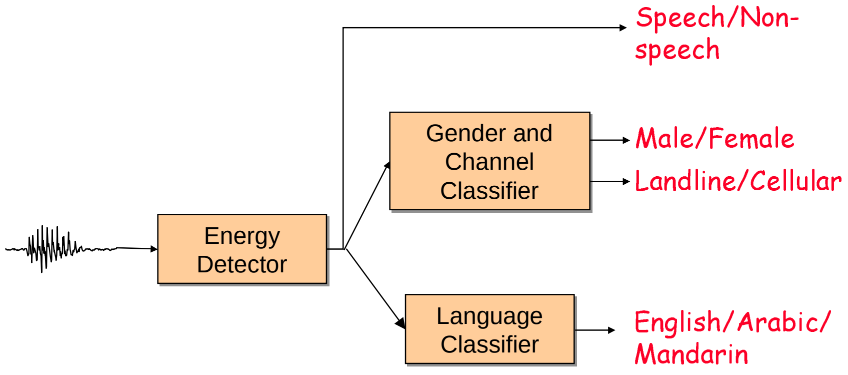


Outline

- CTS Diarization
  - System description
  - Extraction of other metadata
    - Landline vs. Cellular
    - Language identification
  - Analysis of results
- BNEWS Diarization
  - System description
  - Analysis of results
- Conclusions

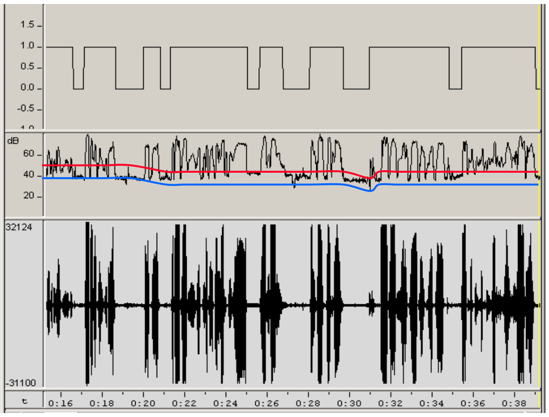
CTS Diarization
System

- In addition to evaluation diarization metadata, system also extracts other metadata



CTS Diarization
Energy Detector

- Adaptive energy based detector
- Detects events with energy above noise floor + threshold
- Fills gaps < 0.3s, removes segments < 0.1s and pads segment boundaries by 0.05s

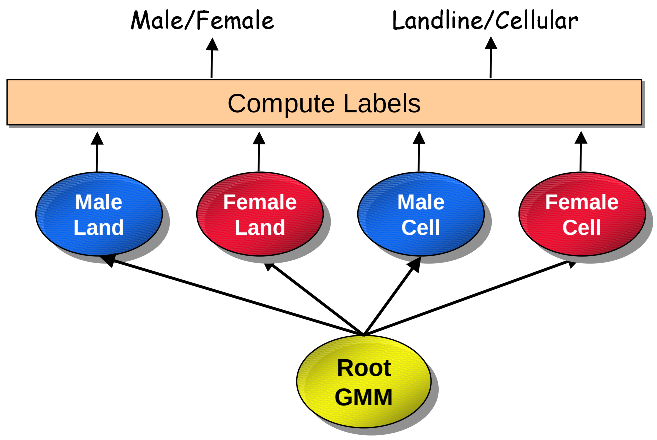


- Search over parameter settings showed no gain in diarization score
- Also found that other speech activity detectors not producing good diarization scores (Talkative, SRSAD)
  - High FA rates
  - Not designed for tight intra-segment marking

CTS Diarization
Gender/Channel Classifier

- Classifier is based on using MAP adapted GMM models
  - Structure from speaker recognition channel compensation work

- Root model is a 2048 GMM trained using pooling of all channel model data
- Channel models are adapted using gender/channel dependent data
- Using adapted GMMs allows use of fast-scoring technique
  - Top-5 mixtures per channel



CTS Diarization
Gender/Channel Classifier

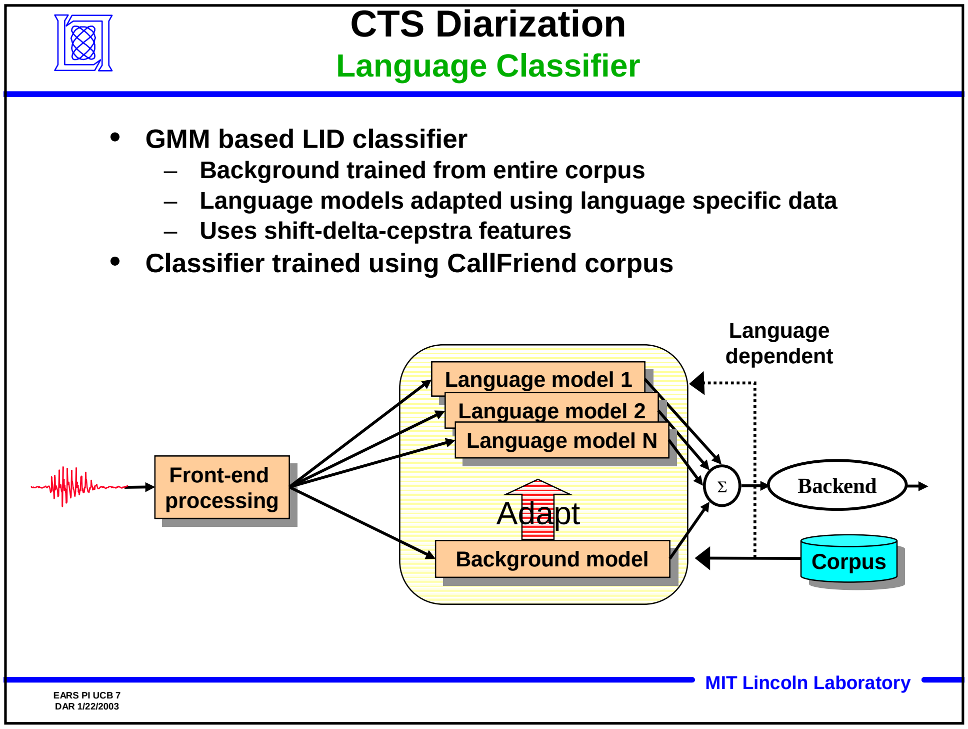
- Classifier has 14 channel models
  - M/F Carb/Elec from Swb2 phase1 (4)
  - M/F from Swb cell part1 (2)
  - M/F Analog/Digital from OGI National Cell (4)
  - M/F from TIMIT [telephone band] (2)
  - M/F from Hub4-96 [telephone band] (2)
- Bayes Classification

$$\Pr(Male \mid X) = \tfrac{1}{M}\sum p(X \mid \text{malemodels}) \Big/ \left(\tfrac{1}{M}\sum p(X \mid \text{malemodels}) + \tfrac{1}{F}\sum p(X \mid \text{femalemodels})\right)$$

$$GID\_label(X) = \begin{cases} Male & if\ \Pr(Male \mid X) > 0.5 \\ Female & if\ \Pr(Male \mid X) < 0.5 \end{cases}$$

$$\Pr(Land \mid X) = \tfrac{1}{L}\sum p(X \mid \text{landmodels}) \Big/ \left(\tfrac{1}{L}\sum p(X \mid \text{landmodels}) + \tfrac{1}{C}\sum p(X \mid \text{cellmodels})\right)$$

$$CID\_label(X) = \begin{cases} Land & if\ \Pr(Land \mid X) > 0.5 \\ Cell & if\ \Pr(Land \mid X) < 0.5 \end{cases}$$



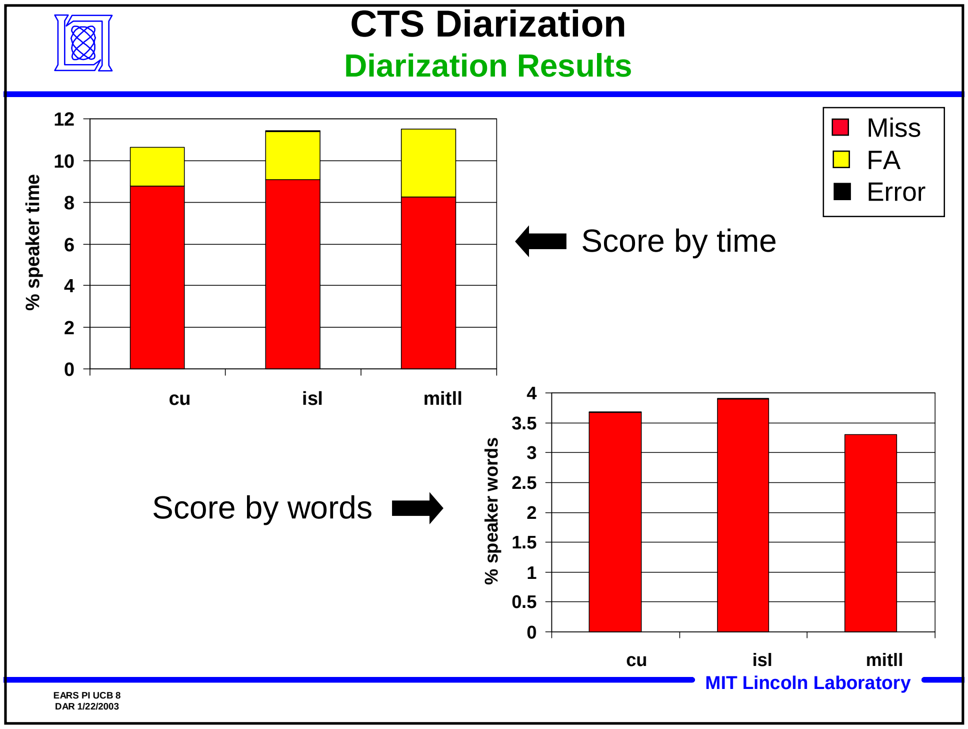
CTS Diarization

Language Classifier

- GMM based LID classifier
  - Background trained from entire corpus
  - Language models adapted using language specific data
  - Uses shift-delta-cepstra features
- Classifier trained using CallFriend corpus

EARS PI UCB 7
DAR 1/22/2003

MIT Lincoln Laboratory



CTS Diarization

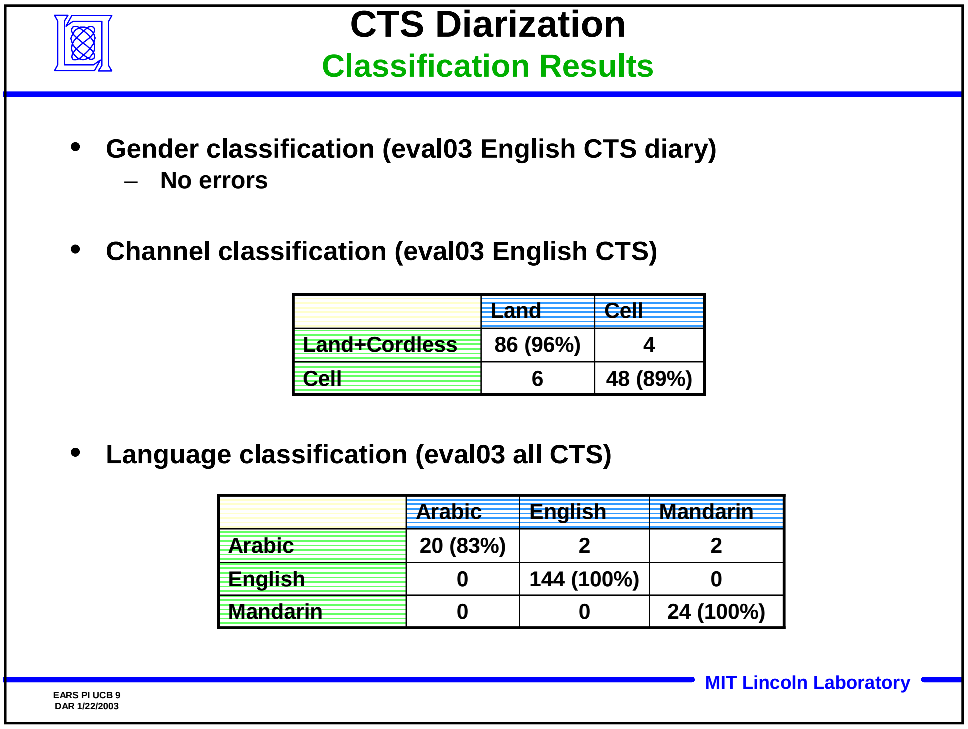
Diarization Results

Score by time

Score by words

EARS PI UCB 8
DAR 1/22/2003

MIT Lincoln Laboratory



CTS Diarization

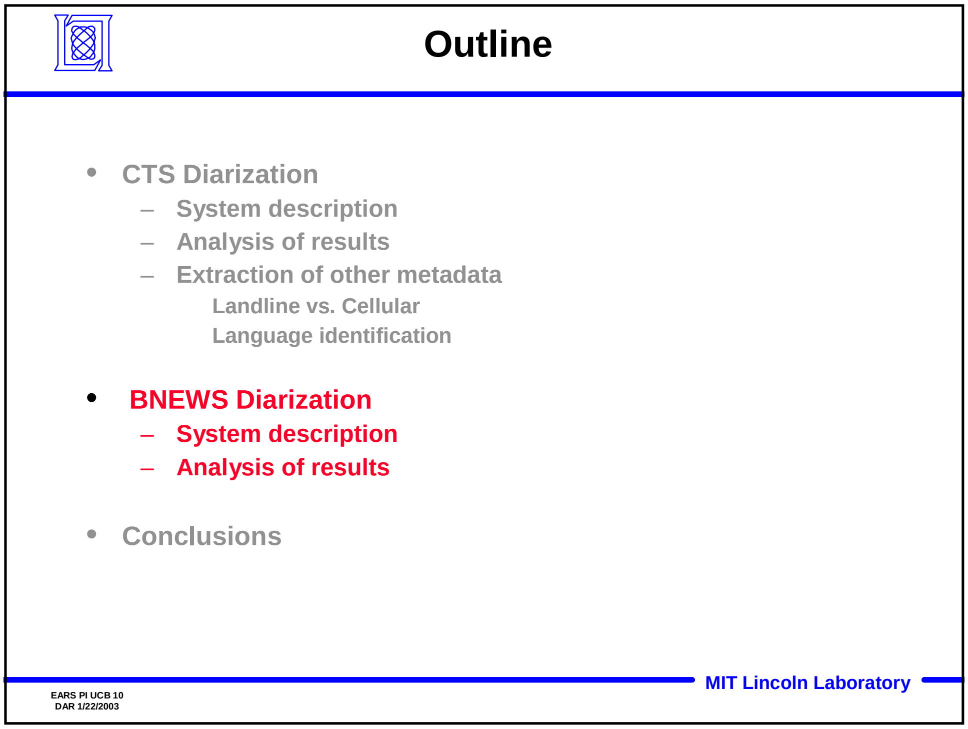
Classification Results

- Gender classification (eval03 English CTS diary)
  - No errors
- Channel classification (eval03 English CTS)

	Land	Cell
Land+Cordless	86 (96%)	4
Cell	6	48 (89%)

- Language classification (eval03 all CTS)

	Arabic	English	Mandarin
Arabic	20 (83%)	2	2
English	0	144 (100%)	0
Mandarin	0	0	24 (100%)



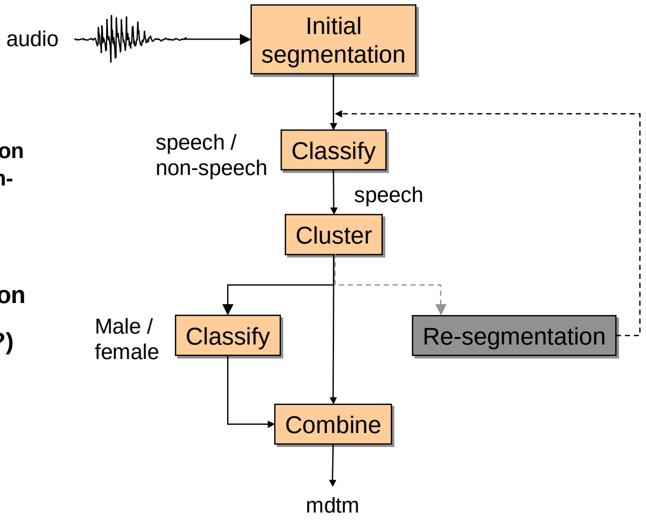
Outline

- CTS Diarization
  - System description
  - Analysis of results
  - Extraction of other metadata
    - Landline vs. Cellular
    - Language identification
- BNEWS Diarization
  - System description
  - Analysis of results
- Conclusions

BNEWS Diarization

Segmentation and Labeling System

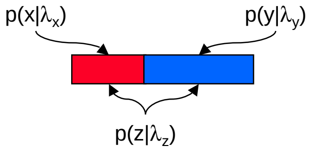
- **Three main components**
  - **Initial segmentation**
  - **Speech/non-speech classifier**
  - **Speaker clustering**
- **Re-segmentation decreased performance (?)**
  - **Not used**



BNEWS Diarization

Change Detection

- **Used BIC based change detection* algorithm**
- **Approach: Search for putative change points using a penalized likelihood ratio test**
  - **Growing search window to find putative change points**
  - **Uses first pass T^2 distance to identify initial change points (Dragon)**



Putative change point if Δ BIC > 0

$$\Delta BIC = -\log \frac{p(z\,|\,\lambda_z)}{p(x\,|\,\lambda_x)\,p(y\,|\,\lambda_y)} - \alpha P$$

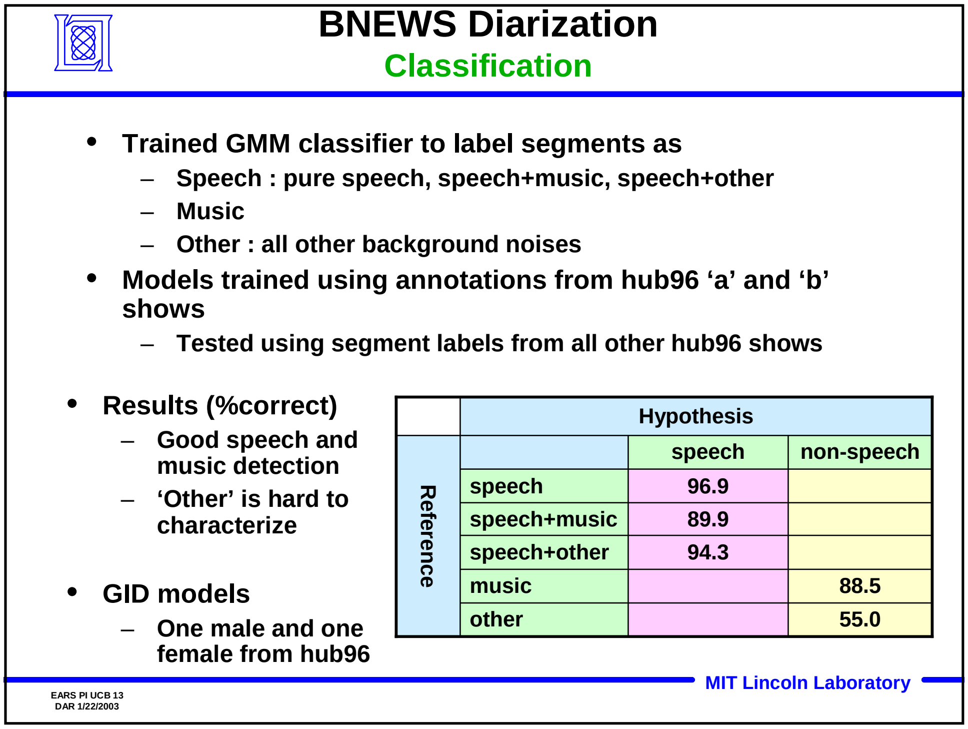
α = BIC weight P = BIC penalty

For full covariance Gaussians

P = 1/2(d+1/2d(d+1)) log N

- **Over segmentation OK since clustering can recombine**
- **Worked very well for BNEWS data**
  - **Best at detecting segment > 2s in duration**
  - **Not very effective for fast interchange (conversational speech)**

"Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", S. Chen and P. Gopalakrishnam, 1998 DARPA Broadcast News Workshop



BNEWS Diarization

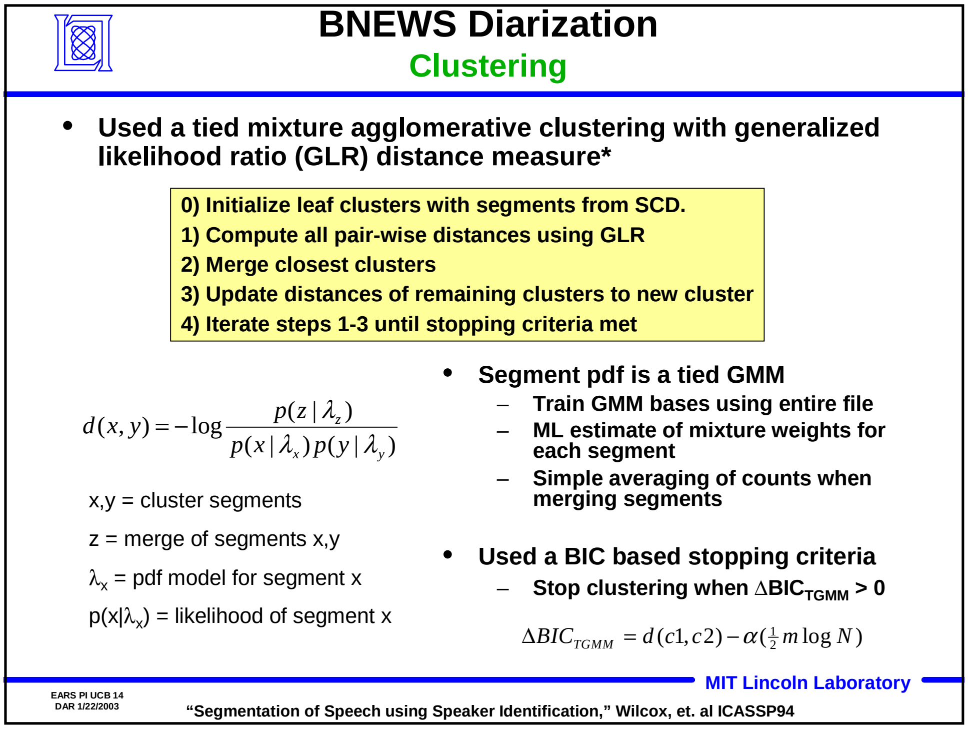
Classification

- Trained GMM classifier to label segments as
  - Speech : pure speech, speech+music, speech+other
  - Music
  - Other : all other background noises
- Models trained using annotations from hub96 ‘a’ and ‘b’ shows
  - Tested using segment labels from all other hub96 shows

- Results (%correct)
  - Good speech and music detection
  - ‘Other’ is hard to characterize

- GID models
  - One male and one female from hub96

		Hypothesis	
		speech	non-speech
Reference	speech	96.9	
	speech+music	89.9	
	speech+other	94.3	
	music		88.5
	other		55.0



BNEWS Diarization

Clustering

- Used a tied mixture agglomerative clustering with generalized likelihood ratio (GLR) distance measure*

0) Initialize leaf clusters with segments from SCD.
1) Compute all pair-wise distances using GLR
2) Merge closest clusters
3) Update distances of remaining clusters to new cluster
4) Iterate steps 1-3 until stopping criteria met

$$d(x,y) = -\log \frac{p(z\,|\,\lambda_z)}{p(x\,|\,\lambda_x)\,p(y\,|\,\lambda_y)}$$

x,y = cluster segments

z = merge of segments x,y

λ_x = pdf model for segment x

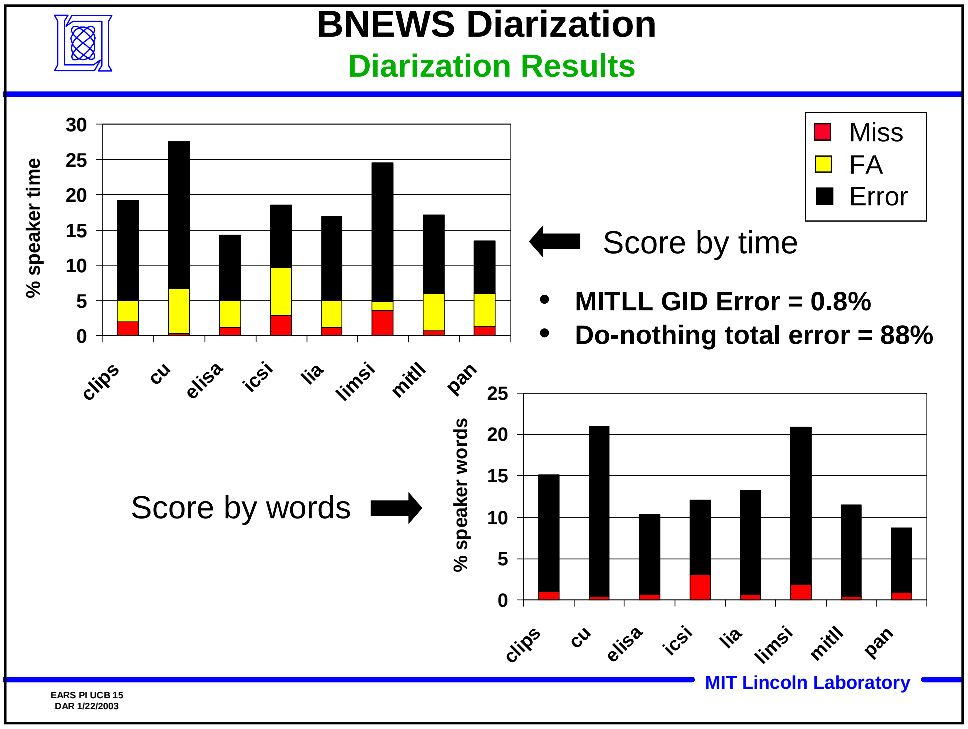
p(x|λ_x) = likelihood of segment x

- Segment pdf is a tied GMM
  - Train GMM bases using entire file
  - ML estimate of mixture weights for each segment
  - Simple averaging of counts when merging segments

- Used a BIC based stopping criteria
  - Stop clustering when $\Delta \mathrm{BIC}_{\mathrm{TGMM}} > 0$

$$\Delta BIC_{TGMM} = d(c1,c2) - \alpha(\tfrac{1}{2} m \log N)$$

“Segmentation of Speech using Speaker Identification,” Wilcox, et. al ICASSP94



BNEWS Diarization

Diarization Results

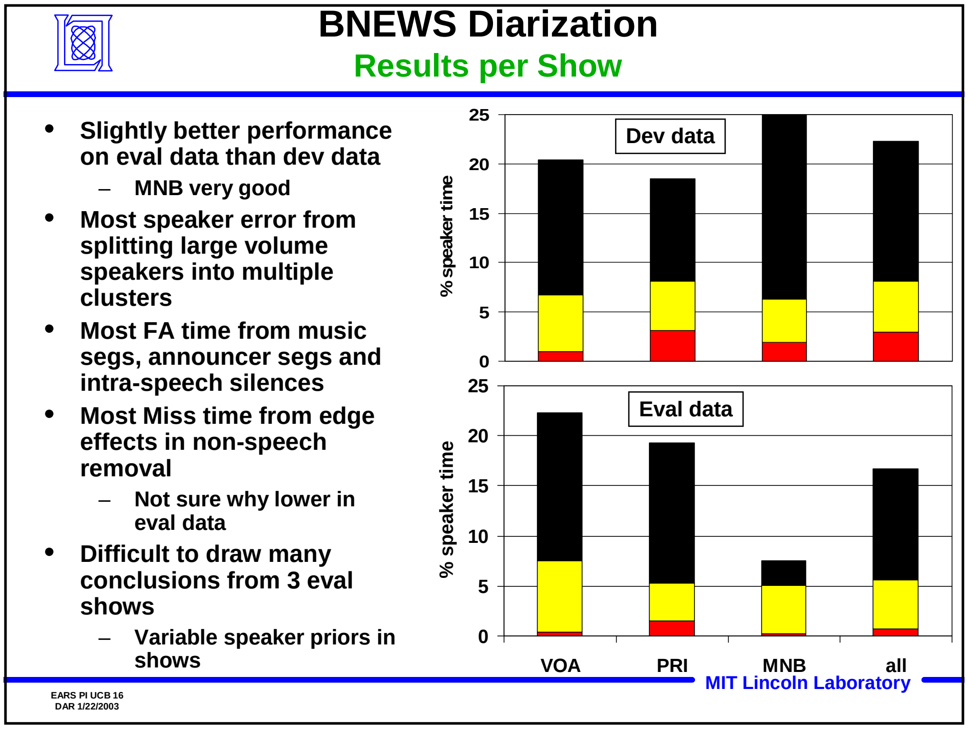
⬅ Score by time

- MITLL GID Error = 0.8%
- Do-nothing total error = 88%

Score by words ➡

MIT Lincoln Laboratory

EARS PI UCB 15
DAR 1/22/2003



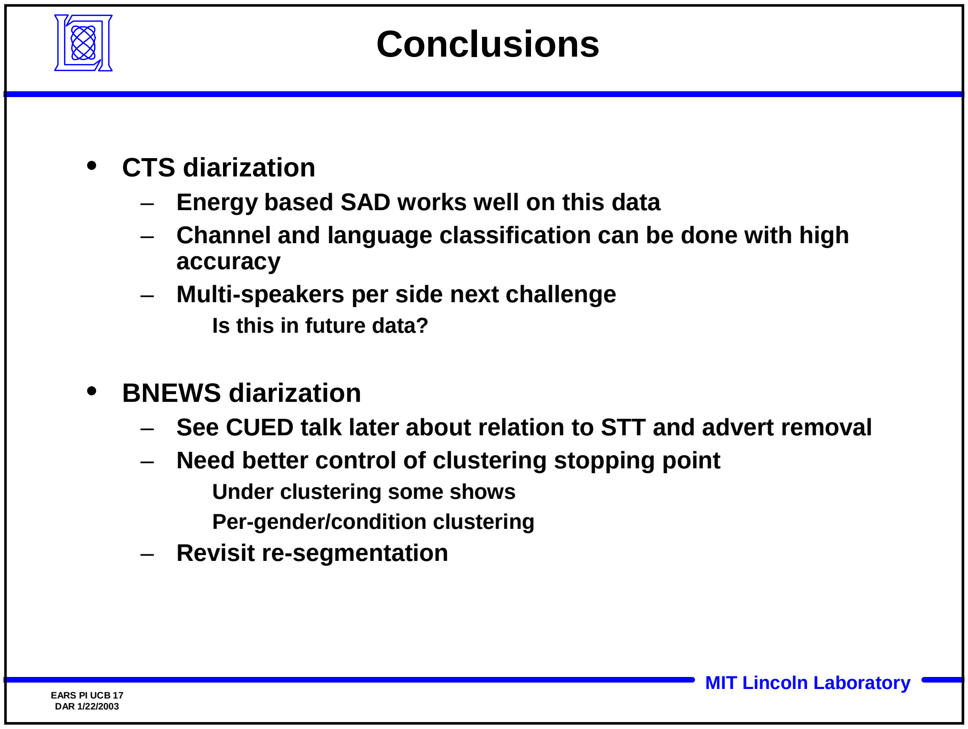
BNEWS Diarization

Results per Show

- Slightly better performance on eval data than dev data
  - MNB very good
- Most speaker error from splitting large volume speakers into multiple clusters
- Most FA time from music segs, announcer segs and intra-speech silences
- Most Miss time from edge effects in non-speech removal
  - Not sure why lower in eval data
- Difficult to draw many conclusions from 3 eval shows
  - Variable speaker priors in shows

MIT Lincoln Laboratory

EARS PI UCB 16
DAR 1/22/2003



Conclusions

- CTS diarization
  - Energy based SAD works well on this data
  - Channel and language classification can be done with high accuracy
  - Multi-speakers per side next challenge
    - Is this in future data?
- BNEWS diarization
  - See CUED talk later about relation to STT and advert removal
  - Need better control of clustering stopping point
    - Under clustering some shows
    - Per-gender/condition clustering
  - Revisit re-segmentation